

Capturing coevolutionary signals in repeat proteins

Rocío Espada[‡] R. Gonzalo Parra[‡] Thierry Mora[§] Aleksandra M. Walczak^{*} Diego Ferreiro^{‡*}

[‡] Protein Physiology Lab, Dep de Química Biológica, Facultad de Ciencias Exactas y Naturales, UBA-CONICET-IQUIBICEN, Buenos Aires, Argentina

[§] Laboratoire de physique statistique, CNRS, UPMC and École normale supérieure, 24 rue Lhomond, 75005 Paris, France

^{*} Laboratoire de physique théorique, CNRS, UPMC and École normale supérieure, 24 rue Lhomond, 75005 Paris, France

^{*} To whom correspondence should be addressed. Email: diegulise@gmail.com

The analysis of correlations of amino acid occurrences in globular proteins has led to the development of statistical tools that can identify native contacts – portions of the chains that come to close distance in folded structural ensembles. Here we introduce a statistical coupling analysis for repeat proteins – natural systems for which the identification of domains remains challenging. We show that the inherent translational symmetry of repeat protein sequences introduces a strong bias in the pair correlations at precisely the length scale of the repeat-unit. Equalizing for this bias reveals true co-evolutionary signals from which local native-contacts can be identified. Importantly, parameter values obtained for all other interactions are not significantly affected by the equalization. We quantify the robustness of the procedure and assign confidence levels to the interactions, identifying the minimum number of sequences needed to extract evolutionary information in several repeat protein families. The overall procedure can be used to reconstruct the interactions at long distances, identifying the characteristics of the strongest couplings in each family, and can be applied to any system that appears translationally symmetric.

Keywords: direct coupling analysis — repeat proteins — direct information — co-evolution

Introduction

The fact that many protein molecules spontaneously collapse stretches of amino acid chains into defined structural domains [Wetlaufer, 1973] facilitates the description, evolution and construction of these peculiar physical objects. Higher order biological *functions* that are correlated with domains can usually be isolated, recombined and adjusted, akin to engineering [Peisajovich and Tawfik, 2007], or tinkering [Jacob, 1977] using modular components. The evolutionary record of natural proteins results from a balance between sequence exploration and constraints: conservation of function within a protein family imposes strong boundaries on sequence variation, sculpting the structural forms visited by members of a protein family. Amino acids that are in spatial proximity in the mean conformational ensemble are expected to co-vary on evolutionary timescales, as their energy contributions to fold stabilization are often localized to groups of residues [Onuchic et al., 1997]. However, correlated residue changes throughout proteins’ history may not necessarily be close in space, as other constraints are always at play [Ferreiro et al., 2007]. Since the evolutionary record is inevitably incomplete, the sequences we find today constitute a biased sample of the possible outcomes, therefore any search for the underlying constraints must take into account contingent factors that may confound the observed correlations. Here we use sequence correlations to explore the link between structure and function in repeat proteins, natural systems for which the identification of functional domains remains challenging [Parra et al., 2013].

Many natural proteins contain tandem repeats of sim-

ilar amino acid stretches. These have been broadly classified in groups according to the length of the minimal repeating units [Kajava, 2012]: short repeats up to five residues usually fold into fibrillar structures such as collagen or silk, while repeats longer than about 60 residues usually fold as independent globular domains. There is a class of proteins whose repeat frequency lies in between these values and for which the folding of the repeating units is coupled. In these periodic repeat proteins unique “domains” are not trivial to define [Parra et al., 2013]. Typical repeat proteins are made up of tandem arrays of ~ 20 -40 similar amino acid stretches that fold into elongated architectures of stacked repeating structural motifs (Fig. 1). Successful design of repeat proteins with novel functions based on simple sequence statistics [Tamaskovic et al., 2012] suggests that folding and functional signals can be partially segregated. Energy landscape theory predicts that foldable polypeptides are much easier to realize in the presence of symmetry as compared to asymmetric arrangements [Wolynes, 1996]. Funneled energy landscapes imply that patterns can form in different parts of the molecule with relative independence and subsequently assemble to higher order structures. This greatly reduces the folding search problem by efficiently arranging relatively small fundamental building blocks in a repetitive fashion [Ferreiro et al., 2008]. Thus, due to the approximate translational symmetry, repeat proteins constitute excellent systems in which to study the coupling between sequential, structural and functional patterns.

The maximum entropy principle proposes a scheme for approaching the problem of extracting essential pair couplings from multiple sequence alignments of families of homologous proteins [Neher, 1994, Weigt et al., 2009,

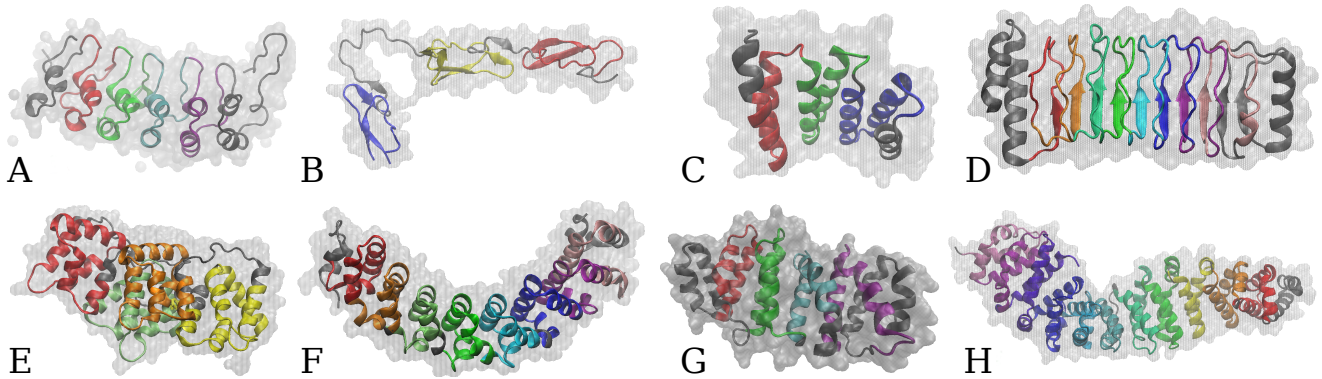


Figure 1: Repeat proteins are formed with tandem arrays of repeats. The crystal structures of members of different repeat protein families are shown, with the backbone colored according to the repeated units. The molecular surface of the repeat array is drawn in transparent gray. **A.** ANK family (PDB:1IKN, chain D), **B.** EGF family (PDB:4D90, chain B), **C.** TPR family (PDB:4GCO), **D.** LEU family (PDB:4NKH, chain A), **E.** ANEX family (PDB:2ZOC, chain A), **F.** PUM family (PDB:2YJY, chain A), **G.** HEAT family (PDB:4G3A, chain A), and **H.** ARM family (PDB:2BCT).

Mora et al., 2010]. The main technical limitations confounding residue correlations are the transitivity of the correlations, the statistical noise due to the relative small number of available observables, and the phylogenetic dependence of the set of sequences assembled into a protein family [Morcos et al., 2014]. Indirect interactions may generate the dominant correlations, and disentangling direct from indirect links is a fundamental step towards inferring the energetics underlying the observed couplings [Weigt et al., 2009]. The application of statistical coupling analysis provides an efficient way of extracting meaningful information from the apparent junk of massive genomic data [Brenner, 1998]. The mean structure of several protein domains can be reasonably well predicted from the statistical analysis of variations in large sets of sequences [Morcos et al., 2011, Sulkowska et al., 2012]. Strong deviations of the statistically coupled positions from the known domain structures leads to explore dynamical aspects of proteins that are related to biological function [Morcos et al., 2013]. Likewise, specific interactions between domains can be characterized and good approximations to the interaction energetics can be obtained [Marks et al., 2011, Cheng et al., 2014, Lui and Tiana, 2013, Mora et al., 2010]). Here we show the limitations of the statistical coupling analysis developed for globular proteins and propose an analogous procedure for quasi-translationally symmetric repeat proteins.

Specifically, we compare the information extracted from two-point correlation functions of multiple alignment of repeat protein sequences to the known structural interactions between the apparent repeated units. We show that the translational symmetry introduces a strong bias in the pair correlations at precisely the length scale of the repeated unit. Equalizing for this bias in an objective way results in correlation matrices from which local native-contacts can be identified. We apply this procedure to many families of repeat protein (introduced in Fig. 1) and show that some families have strong interactions mainly between repeats, while others mainly within

single repeats. These observations can be linked to the functional characteristics of the families.

Results

Direct coupling analysis of repeat proteins

To characterize correlations between amino acid positions in natural repeat proteins we needed to define a length scale on which to search, align and compose a repeat region. We chose to use the minimal definition of repeats present in the PFAM database [Bateman et al., 2004], and obtained sequences of single repeated units for the families listed in Table 1 of SI. Since a repeat domain is formed with multiple tandem copies of repeats units [Parra et al., 2013], the minimal sequence that includes an interface between repeats is composed of two consecutive units. We thus constructed multiple sequence alignments (MSA) of pairs of consecutive repeats for each family. The sets of sequences were corrected for phylogenetic bias and finite-size sampling as described in the Methods.

Mutual information (MI) and direct information (DI) use covariance in homologous protein sequences to deduce structural constraints. While MI uses the joint frequencies of aminoacids (eq. 1), DI (eq 2) uncouples direct interactions from interactions mediated by a third residue on the complete sequence of the protein. The upper triangles of figures 2B and 2C show the MI and DI matrices for one of the most abundant repeat proteins, the Ankyrin-repeat family. The typical length of these repeats is 33 residues, so values on columns/rows 1 to 33 and 34 to 66 correspond to interactions between residues within a repeat, while values on columns 1 to 33 and rows 34 to 66 correspond to interactions between residues on consecutive repeats. Both MI and DI present overall similar patterns with MI having a noisier background signal. The values corresponding to pairs of positions on consecutive repeats reach comparable values to those within each repeated unit. There appears to be as much evolutionary correlations between residues on the same repeat as between residues in consecutive repeats. A question that arises is whether the strong signal

between repeats is due to the inevitable similarity of the sequences of repeat regions or to true coevolutionary interactions between neighboring repeats.

A close inspection of the couplings detected between repeated units reveals that the strongest signals are attributed to pairs of positions that are 33 residues apart (Fig. 2B and 2C, upper triangle). Since the ankyrin repeats aligned are of this precise length L_0 , these apparent interactions occur between residues that occupy equivalent positions in each repeat, i.e.: the pair of positions $(i, i + L_0)$ corresponds to the i th residue on the first repeat and the i th residue on the second repeat. If repeats in proteins were identical, the interactions between residue i and $i + L_0$ should get maximum MI and DI values as these would show perfect co-variation. At the same time, the submatrix of positions between repeats should be identical to the submatrix of pairs of positions within the repeats. Thus, the identity between repeated units should be taken into account when evaluating correlations between repeats.

To characterize how the identity between neighboring repeats affects the covariation analysis, we compared the distribution of the percentage of identical residues, $\%Id$, between pairs of consecutive repeats, and between randomly assembled pairs of repeats (Fig. 2). For the ankyrin family, the distribution of $\%Id$ for random pairs is a Gaussian centered around 30%, while the natural pairs show higher mean and a large tail towards higher $\%Id$ values (Fig. 2A). This higher similarity between pairs of consecutive repeats is expected to induce correlations between i and $i + L_0$ positions, as observed. To compensate for the higher $\%Id$ between natural repeats we developed a correction factor that equalizes the effects of quasi-translational symmetry. This correction consists of calibrating the weight of each sequence in the natural neighbors according to the $\%Id$ between the component pair of repeats, and rescaling it so that it matches the expected frequency of $\%Id$ between random pairs of repeats of the same family (see Methods). We refer to the obtained values as DI_{id} and MI_{id} . Figures 2B and 2C show the DI and MI corrected only for phylogeny and finite counts (upper triangles), together with the ones that include this additional factor DI_{id} and MI_{id} (lower triangles). The strong symmetric $(i, i + 33)$ off-diagonal signal is attenuated for both MI and DI, as expected if the signal originates from biases in the $\%Id$ distributions. Importantly, the DI values obtained for interactions between all other positions are not significantly affected by the $\%Id$ equalization.

The same analysis was performed for all the other repeat protein families (see Suppl. material, Fig. S1). The results for the TPR, EGF and LEUCINE RICH families show a strong bias in the symmetric $(i, i + L_0)$ interactions. These also show a higher sequence identity between true first neighboring repeats, which biases the inter-repeat couplings. Families ARM, SPECTRIN, ANNEXIN and PUMILIO do not show a high $(i, i + L_0)$ signal on the DI and MI matrices. For these, the distributions of $\%Id$ between true and random neighbors are similar, consistent with the notion that the symmetric signal is caused just by the bias in similarity between neighboring repeats. Applying the $\%Id$

equalization to these families does not significantly change the DI and MI values, showing that the correction is not detrimental to the overall procedure. Finally, the NEBULIN family does not present a strong $(i, i + L_0)$ signal, and the HEAT family has a very rugged $\%Id$ distribution. We believe that these effects are caused by an insufficient number of effective sequences on the alignments, which cannot ensure a robust calculation of DI and MI (*vide infra*). We conclude that sequences of proteins that show quasi-translational symmetry should be treated with an additional correction factor to account for the biases that the internal sequence identity can bring about.

Prediction of native contacts for repeat proteins

For several globular domains it has been shown that native contacts can be inferred from the inspection of the top-list of residue pairs according to the DI ranking [Morcos et al., 2014]. There is no established way to discern the minimum value of DI to be used as the cutoff, as these depend on the topology of the fold, the sampling of sequences and the details of the method used to obtain DI, thus 50 to 200 pairs are empirically used. Since domains of repeat proteins are composed with multiple copies of repeated units, we asked whether DI and DI_{id} metrics are useful predictors of direct native interactions at the sub-domain level. We observed that the absolute values of DI we calculated for pairs of repeats are lower than those computed for globular domains, (Fig. 2 and S1), complicating the distinction of positive DI outliers from the background signal. We developed a clustering method to objectively delimit the true positive interactions. We first calculated the euclidean distance between each pair of DI values as $dDI_{a,b} = \sqrt{(DI_a - DI_b)^2}$; and made a hierarchical clustering of the obtained distances. To delimit the clusters we used the dynamic tree cut method [Langfelder et al., 2008], which allows us to distinguish nested clusters. We found that most of the DI pairs fall in one big cluster which we assigned to the background signal (Fig. S2). The other clusters have fewer members and constitute outliers of the normal distribution. We consider the true coevolutionary signals as those within small clusters of positive DI values.

Several high-resolution structures for repeat proteins are available. These typically fold into elongated architectures where most members of a family display an overall similar topology (Fig. 1). Notably, the repeat arrays can vary in the number of repeat units and many details and irregularities plague the structural representatives [Parra et al., 2013]. To get an overall representation of the distribution of contacts in the known structures we computed the probability of contact formation along the ensembles of structures as described in Methods. We obtained an average contact map of the repeat architecture mapped on to the sequences of the MSA, where residue pairs with high density correspond to residue pairs that are most frequently encountered within contact distance (Fig. 3, lower triangle). The pattern of evolutionary interactions inferred from the clustering of DI_{id} is remarkably similar to the experimental contact map densities for most

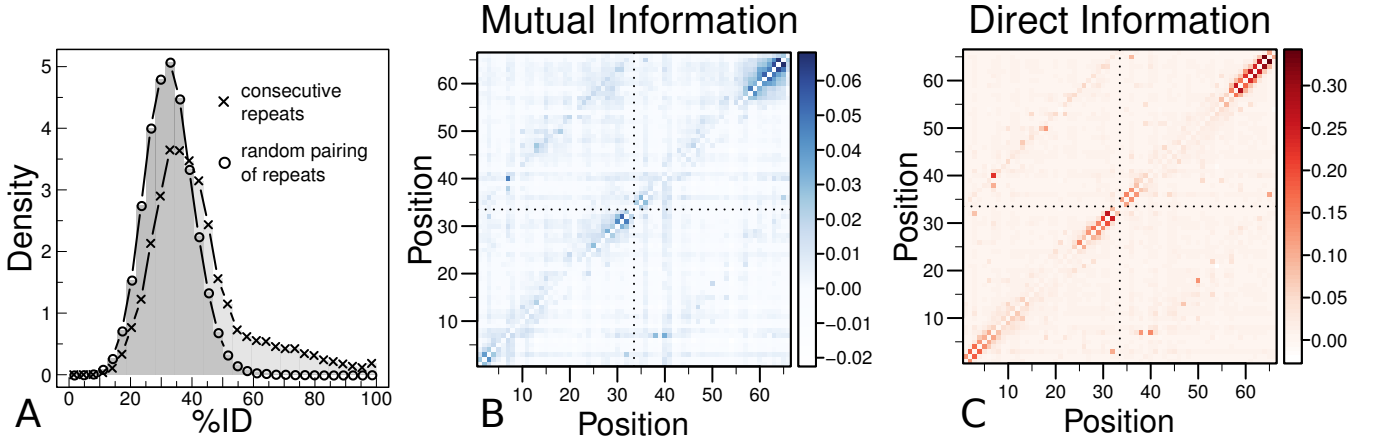


Figure 2: The sequence identity between repeated units can bias the inference of evolutionary couplings. Repeat sequences of the ANK family were concatenated in a MSA of size $2L_0 = 66$ positions and ≈ 73000 sequences and co-variations were measured with mutual and direct information metrics. **A.** Sequence identity distributions between consecutive ANK repeats found in (x) natural proteins and (o) randomized pairs of repeats. **B.** Mutual information and **C.** Direct information matrices between positions obtained without correcting (upper half) or with proper equalization for repeat identity (lower half).

families (Fig. 3 and S2). The signals from the pairs of positions of consecutive repeats ($i, i + L_0$) do not always correspond to a high contact probability, yet if present they are confidently detected.

One of the longest pairs of repeated units we study belongs to the Annexin family ($2L_0 \approx 132$ residues). The DI_{id} hits strongly resemble the average contact map, with 113 out of the 150 DI_{id} pairs found within contact distance in at least half of the experimental structures (Fig 3A). Most of these correspond to interactions within each repeat, with few interactions at the repeat interfaces, unlike the correlations found in other repeat proteins, such as the Ankyrin family (Fig. 3B)

The clustering procedure assigns 246 hits for DI_{id} , 166 of which are typically found within contact distance. Most of these are found outside the usual binding site of these proteins – the β -hairpin motif. Coevolutionary interactions of the Pumilio family also map to positions between repeated units, with 113 experimental contacts out of 150 predicted (Fig. 3C). Yet in this case they are mainly clustered around the regions where these proteins bind nucleic acids. Within the top 237 DI_{id} identified for the Tetratricopeptide family, only 167 are typically found within contact distance in the experimental structures and most of the outliers are in regions physically compatible with the known structures (Fig. 3D). A similar picture is apparent in the Armadillo family, where only 140 interactions correspond to mean contacts among the 231 predicted (Fig. S2). In the case of the Leucine-rich family, few interactions appear as outliers in DI_{id} distribution, and most of them have been observed to form close contacts between repeated units (Fig. S2). Repeats of the EGF family rarely interact, and DI_{id} consistently fails to detect inter-repeat correlations, acting as a negative control for the overall procedure (Fig. S2). Finally, few co-evolutionary interactions are assigned in the HEAT family, probably due to the limited number of available sequences (see below). Since

there are no experimental structures for the Nebulin family, we cannot evaluate if the identified DI_{id} hits correspond with native-contacts.

Distant couplings along a repeat array

Folding of repeat domains usually involves the cooperative formation of structures at a length scale that exceeds first neighbors [Aksel and Barrick, 2009]. Folding in some regions nucleates the folding of contiguous segments, allowing for a quasi-one-dimensional treatment of the dynamics [Ferreiro and Wolynes, 2008]. A natural question that arises is how do evolutionary couplings in and between repeats change as the separation between the repeats increases.

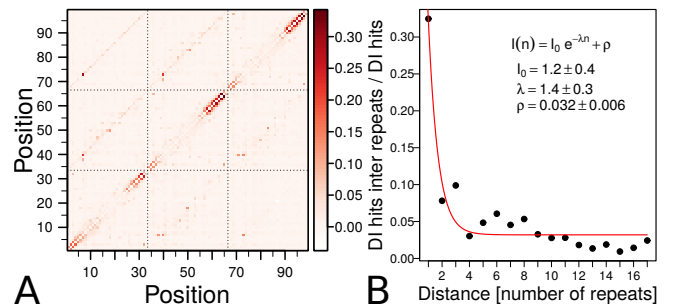


Figure 4: Correlations along ANK repeat arrays. **A.** Direct information matrix calculated for three pairs of ANK repeats without (upper triangle) or with (lower triangle) the DI_{id} equalization. **B.** Proportion of DI_{id} hits between repeated units for alignments of n -th neighbors. The line is a non-linear fit of the data to an exponential decay.

An analogous correction to the weights of the sequences must be made to treat n -neighbors interactions (see lower triangle of Fig. S3 for the uncorrected DCA of three consecutive repeats of the ankyrin family). When

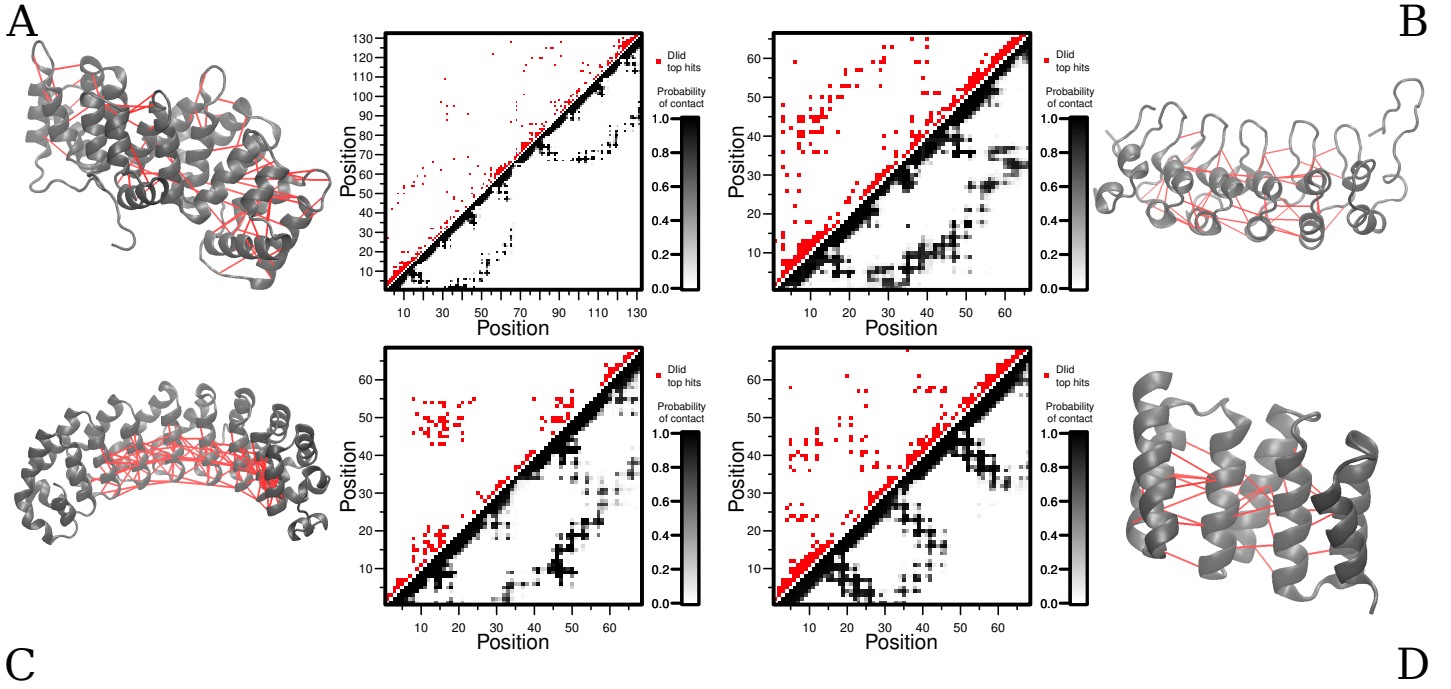


Figure 3: Native contacts can be predicted from the identity-equalized direct information DI_{id} . The probability of contact formation along ensembles of structures of consecutive repeat pairs in several repeat protein families is shown (lower triangles), together with the contact prediction based on the DI_{id} distribution (upper triangles). The structure of representative family members are shown on the sides, with the backbones as gray ribbons and the first 20 predicted contacts along multiple repeat pairs in red. **A.** ANEX (PDB:2ZOC, chain A) **B.** ANK (PDB:11KN, chain D) **C.** PUM (PDB:2YJY, chain A) **D.** TPR (PDB:4GCO)

the proper equalization is performed, the symmetric signals attenuate and the true coevolutionary correlations appear (DI_{id} lower triangle of Fig. S3). In principle the correction to the symmetric $(i, i + nL_0)$ interactions can be applied to arbitrarily large repeat proteins. Yet the sampling needed is much larger and the computing time grows as L^2 , restricting the application to longer repeat arrays. Since in ANKs, as in most of the repeat protein families, interactions are concentrated at relatively short sequences separations, we reconstructed a DI_{id} matrix from a parallel calculation of repeat pairs. For first neighbors we estimated DI_{id} as described previously, and for second neighbors we concatenated the sequences in an MSA of size $2L_0$. The reconstructed matrix for all interactions is very similar to the one calculated on the whole three-repeat MSA (Fig. 4A and S3), facilitating the application of the analysis for larger repeat arrays.

We observed that as the separation between repeats increases, the DI_{id} between repeats decays significantly (Fig. 4B). True repeat pair interactions are less frequent, and this is reflected in the evolutionary couplings between units. The number of interactions between repeats decreases roughly exponentially with repeat separation, with a half-length of about 1.4 repeats (Fig. 4B), suggesting that the evolutionary correlation length of Ankyrin repeat arrays is ~ 1.5 units.

Robustness and confidence of the analysis

For a robust calculation of the DI one must have a sufficiently large number of effective sequences to approximate the marginal and joint probability distributions from

the observed frequencies of occurrences of amino acids. Since there is no general principle indicating how many sequences are necessary and sufficient for robust estimation, we empirically quantified the minimum number of effective sequences in various repeat protein families.

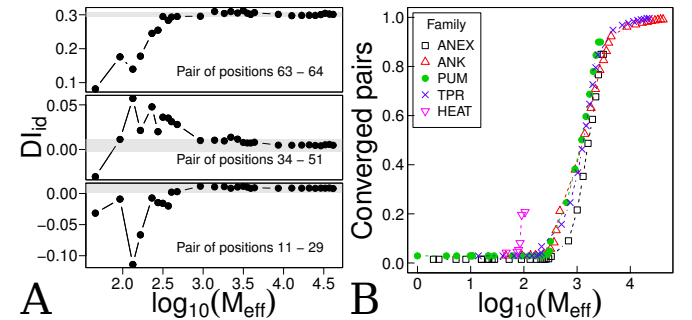


Figure 5: Robustness of the DI_{id} procedure. Subsets of alignments were constructed by recurrently removing random groups of sequences from each dataset of repeat pairs. M_{eff} is the number of effective sequences used in the alignment. **A.** Particular examples of the stability of DI_{id} assignments as sampling changes on the ANK family. The gray shadow delimits the 1% fluctuation interval set as a convergence criteria. **B.** Overall stability of the DI_{id} assignments in several repeat protein families.

We constructed subsets of alignments by recurrently removing random groups of sequences from each dataset of repeat pairs, and calculated DCA on each of these subsets. The reduction in the number of sequences typically de-

crease the absolute values of the high ranking DI_{id} matrix elements and at the same time increases the background DI_{id} signals (Fig. S4), making both MI_{id} and DI_{id} signals indistinguishable from the background for small sample sizes.

For well determined parameters we expect the true value will be better estimated as sampling increases. Examples of the robustness of the DI_{id} assignments are shown in the panels of Fig. 5A. While the DI_{id} of some residue pairs can be confidently established with about 500 effective sequences, other pairs do not reach stable values even when all the available sequences are taken into account (Fig. 5A). To globally quantify the convergence of the DI_{id} matrix we evaluated how many of the residue pairs reach a limiting value within 1% of the one obtained with the largest sample size. For every subset of sequences, s , we require that $|DI_{ij}^s - DI_{ij}| < 0.01 \cdot (\max(DI) - \min(DI))$, where DI_{ij}^s is the DI between position i and j calculated over the s -th subset, DI_{ij} is the DI on the largest set of sequences, and $\max(DI)$ and $\min(DI)$ are the maximum and minimum values for all positions in all subsets. Additionally all subsets larger than the subset s one must have a standard deviation lower than 1% of the standard deviation of the DI values from all the subsets. If a residue pair fulfills these conditions, we say it has converged at the particular s sample size. We quantified how many of the residue pairs satisfy the convergence criteria at various sample sizes (Fig. 5B). The best sampled families, ANK and TPR, contain enough sequences to converge the DI_{id} for almost all residue pairs of consecutive repeats. Reducing the number of input sequences results in a loss of convergence of some sites; the DI_{id} of around 90% of the residue pairs can be confidently established with about 10% of the total sequences ($M_{eff} \approx 10^{7/2}$) (Fig. 5B). If the subsamples are further reduced, the proportion of positions that converge drops catastrophically. Yet even more relaxed criteria for convergence give confident results for the high-ranking DI pairs, as exemplified by the PUM and ANEX families (Fig. 5B). However the samples for the HEAT family are not sufficient to confidently quantify repeat pairs co-evolution.

Discussion

Repeat proteins are formed with various tandem repetitions of similar amino acid stretches. Due to the approximate translational symmetry, regions in proximity in the amino acid chain show similarities in their sequence patterns, which can result in close to perfect co-variation in a multiple sequence alignment and hence bias the inferred interactions between residues (Fig. 2). To compensate for this natural bias we developed an equalization that re-weights each sequence in the multiple alignment to account for correlations characteristic of the protein family. This procedure reveals the true co-evolutionary signals in the case of strong biases, importantly leaving the quantifications unchanged in the absence of bias.

The DI_{id} metric resulting from this corrected statistical coupling analysis is a good predictor of native interactions at the sub-domain level for proteins with a quasi transla-

tional symmetry, similarly to the original DI metric for globular proteins [Morcos et al., 2014]. The highest ranking DI_{id} pairs are usually found in spatial proximity in all of the repeat protein families analyzed (Figs. 3 and S2). Interestingly, the patterns of co-evolutionary interactions are not a random subset of all the native-interactions, but segregate into particular groups in each family. Some families display relative high inter-repeat correlations, while in others the repeats appear to be independent evolutionary units. In their native environment, most repeat proteins participate in binding other macromolecules, and are thus expected to show co-variations in the positions that correspond to the binding interface. We observed that some architectures do show higher co-variations at the typical binding interface, like the nucleic-acid binding PUM family, while in the ubiquitous ANK family the typical binding interface is depleted of DI_{id} pairs.

A reliable estimation of DI requires a sufficiently large number of sequences. This number depends on the length, the topology and the ontology of the proteins under scrutiny. We empirically quantified the minimum number of effective sequences needed by sampling the subsamples of repeat protein families (Fig. 5). In most families we found that $\sim 90\%$ of the residue pairs can be confidently established with $\sim 10^{7/2}$ sequences (Fig. 5). The highest ranking DI interactions confidently predict native contacts even for much scarcer sampling.

Repeat proteins usually fold cooperatively several consecutive repeats [Aksel and Barrick, 2009]. Nucleation of the folding in some region facilitates the folding of contiguous segments, allowing for a quasi-one-dimensional treatment of the dynamics [Ferreiro and Wolynes, 2008]. We found that the statistical couplings calculated from sequence variations in the ANK family decay roughly exponentially (Fig. 4) as the separation between the repeats increases. The predicted global correlation length of ~ 1.4 repeated units is remarkably close to that inferred from statistical mechanical analysis of folding experiments [Street and Barrick, 2009, Wetzel et al., 2008] and folding simulations [Ferreiro et al., 2005]. These predictions are based on approximating long-range covariations from sets of pair-wise inter-repeat interactions, allowing for the application of the procedure for arbitrarily large structures for which an exact calculation would be computationally prohibitive.

Materials and methods

Alignments data We obtained the MSA for repeat units with NCBI data from the PFAM [Finn et al., 2013] database for the families listed on Table 1 of SI. For each MSA we ignored the columns that contain gaps in more than the 80% of the members. The remaining number of residues in each case is referred as L_0 . In order to reconstruct tandem arrays of repeats, we concatenated the sequences that belong to the same protein (as identified in Uniprot [Consortium, 2014]), and for which the sequence separation is less than $L_0/3$. The alignment thus generated is referred as first neighbor alignment and has $L = 2L_0$ columns (positions) with M rows (sequences) for each of the prototypical families of repeat proteins listed in Table 1 of SI.

DCA calculations On every constructed MSA we performed DCA using the matrix inversion method detailed in [Morcos et al., 2011]. To correct for the phylogenetic bias in the ensembles of sequences, we weighted them with the Henikoff and Henikoff heuristic [Henikoff and Henikoff, 1994], by assigning a weight $w_i = \sum_j \frac{1}{r_j \cdot s_j^i}$ to each sequence. r_j is the number of different amino acids present in position j of the MSA and s_j^i is the number of sequences that have the same amino acid on position j than sequence i . We approximated the effective number of sequences as $M_{eff} = \sum_i w_i$. We calculated the mutual information (MI) and direct information (DI) as:

$$MI_{ij} = \sum_{A,B} f_{ij}(A,B) \ln \left(\frac{f_{ij}(A,B)}{f_i(A)f_j(B)} \right) \quad (1)$$

$$DI_{ij} = \sum_{A,B} P_{ij}^{dir}(A,B) \ln \left(\frac{P_{ij}^{dir}(A,B)}{f_i(A)f_j(B)} \right) \quad (2)$$

where $f_i(A)$ is the marginal frequency of amino acid A at position i of the MSA, $f_j(B)$ is the marginal frequency of amino acid B at position j of the MSA, $f_{ij}(A,B)$ is the joint frequency of having amino acid A at position i and amino acid B at position j simultaneously and $P_{ij}^{dir}(A,B)$ is the probability of having amino acid A at position i and amino acid B at position j simultaneously generated by the direct coupling between these pairs of residues.

The finite-size of the ensemble of sequences generates spurious correlations that must be corrected for. By scrambling each of the columns of a natural MSA we generate MSA_{IM} which keeps the marginal frequencies of the amino acids in each position but breaks all true correlations. We calculated mutual and direct information for this site-independent alignment and subtracted the results from the mutual and direct information calculated on the original MSA. These values are presented in the matrices MI (for mutual information) and DI (for direct information).

DI_{id} calculation We accounted for the self-similarity of repeats by weighting the sequences according to the sequence identity of a repeat pair. We calculated the percentage of identical residues %Id between the repeats on the same sequence ($\nu(\%Id)$) and the randomly expected %Id between M pairs of repeats of the same family, but belonging to different proteins, $\nu^{random}(\%Id)$. Since aligned repeats have L_0 residues each, the %Id can only take discrete values n/L_0 with n an integer between 0 and L_0 . We weighted each sequence by:

$$w_i^c = w_i \frac{\nu^{random}(\%Id = n_0 L_0)}{\nu(\%Id = n_0 L_0)} \quad (3)$$

where w_i is the Henikoff weight of a sequence that has %Id = $n_0 L_0$. The DCA calculations that include these weights are referred to as DI_{id} and MI_{id}.

Density of contacts map To compare the results of DI and MI calculations with available structural models of repeat proteins we took all available structures from the protein data bank [Bernstein et al., 1977] cataloged under the PFAM accession number of the family. The numeration of the residues for the repeat units were identified with HMMER [Finn et al., 2011] and the corresponding MSA. We calculated a contact map for each PDB structure based on the euclidean distances between the C α atoms of each amino acids. We considered amino acids to be in contact if their C α are closer than 10 Å. The contact probability of a pair of residues was defined

as the number of times this pair is found in contact in the ensemble of structures.

Acknowledgments Work was supported by the Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina (CONICET), the Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT), and ERCStG n. 306312. .

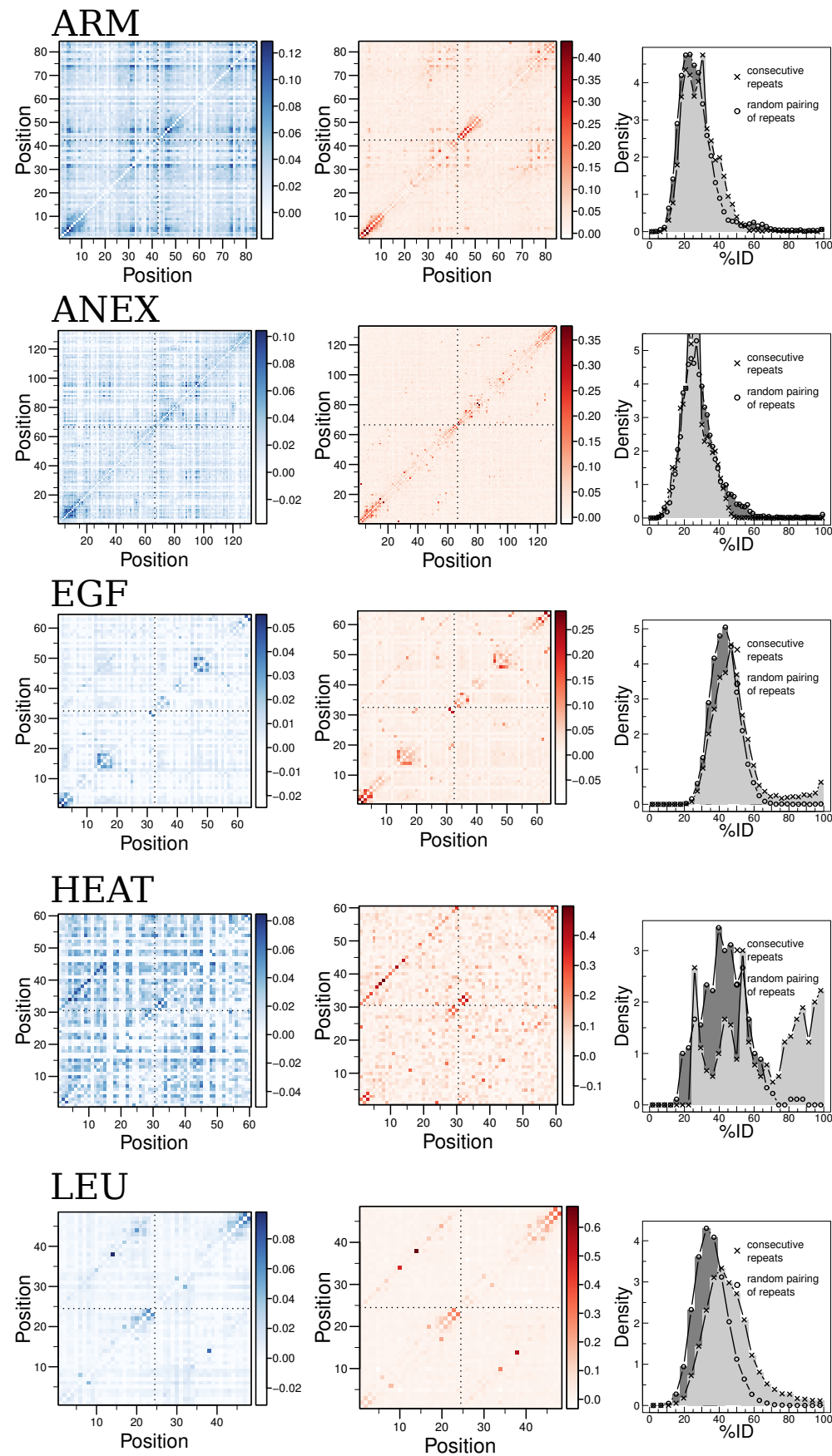
References

- [Aksel and Barrick, 2009] Aksel, T. and Barrick, D. (2009). Analysis of repeat-protein folding using nearest-neighbor statistical mechanical models. *Methods in enzymology*, 455:95–125.
- [Bateman et al., 2004] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., et al. (2004). The pfam protein families database. *Nucleic acids research*, 32(suppl 1):D138–D141.
- [Bernstein et al., 1977] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The protein data bank. *European Journal of Biochemistry*, 80(2):319–324.
- [Brenner, 1998] Brenner, S. (1998). Net prophets. *Curr. Biol.*, 8(5):R147.
- [Cheng et al., 2014] Cheng, R. R., Morcos, F., Levine, H., and Onuchic, J. N. (2014). Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc. Natl. Acad. Sci. U.S.A.*, 111(5):E563–571.
- [Consortium, 2014] Consortium, T. U. (2014). Activities at the universal protein resource (uniprot). *Nucleic Acids Research*, 42(D1):D191–D198.
- [Ferreiro et al., 2005] Ferreiro, D. U., Cho, S. S., Komives, E. A., and Wolynes, P. G. (2005). The energy landscape of modular repeat proteins: topology determines folding mechanism in the ankyrin family. *Journal of molecular biology*, 354(3):679–692.
- [Ferreiro et al., 2007] Ferreiro, D. U., Hegler, J. A., Komives, E. A., and Wolynes, P. G. (2007). Localizing frustration in native proteins and protein assemblies. *Proc. Natl. Acad. Sci. U.S.A.*, 104(50):19819–19824.
- [Ferreiro et al., 2008] Ferreiro, D. U., Walczak, A. M., Komives, E. A., and Wolynes, P. G. (2008). The energy landscapes of repeat-containing proteins: topology, cooperativity, and the folding funnels of one-dimensional architectures. *PLoS computational biology*, 4(5):e1000070.
- [Ferreiro and Wolynes, 2008] Ferreiro, D. U. and Wolynes, P. G. (2008). The capillarity picture and the kinetics of one-dimensional protein folding. *Proceedings of the National Academy of Sciences*, 105(29):9853–9854.
- [Finn et al., 2013] Finn, R. D., Bateman, A., Clements, J., Coghill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2013). Pfam: the protein families database. *Nucleic acids research*, page gkt1223.
- [Finn et al., 2011] Finn, R. D., Clements, J., and Eddy, S. R. (2011). Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, page gkr367.

- [Henikoff and Henikoff, 1994] Henikoff, S. and Henikoff, J. G. (1994). Position-based sequence weights. *Journal of molecular biology*, 243(4):574–578.
- [Jacob, 1977] Jacob, F. (1977). Evolution and tinkering. *Science*, 196(4295):1161–1166.
- [Kajava, 2012] Kajava, A. V. (2012). Tandem repeats in proteins: from sequence to structure. *Journal of structural biology*, 179(3):279–288.
- [Langfelder et al., 2008] Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*, 24(5):719–720.
- [Lui and Tiana, 2013] Lui, S. and Tiana, G. (2013). The network of stabilizing contacts in proteins studied by coevolutionary data. *J Chem Phys*, 139(15):155103.
- [Marks et al., 2011] Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3d structure computed from evolutionary sequence variation. *PloS one*, 6(12):e28766.
- [Mora et al., 2010] Mora, T., Walczak, A. M., Bialek, W., and Callan, C. G. (2010). Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences*, 107(12):5405–5410.
- [Morcos et al., 2014] Morcos, F., Hwa, T., Onuchic, J. N., and Weigt, M. (2014). Direct coupling analysis for protein contact prediction. *Methods Mol. Biol.*, 1137:55–70.
- [Morcos et al., 2013] Morcos, F., Jana, B., Hwa, T., and Onuchic, J. N. (2013). Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc. Natl. Acad. Sci. U.S.A.*, 110(51):20533–20538.
- [Morcos et al., 2011] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.*, 108(49):E1293–1301.
- [Neher, 1994] Neher, E. (1994). How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences*, 91(1):98–102.
- [Onuchic et al., 1997] Onuchic, J. N., Luthey-Schulten, Z., and Wolynes, P. G. (1997). Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem*, 48:545–600.
- [Parra et al., 2013] Parra, R. G., Espada, R., Sánchez, I. E., Sippl, M. J., and Ferreiro, D. U. (2013). Detecting repetitions and periodicities in proteins by tiling the structural space. *The Journal of Physical Chemistry B*, 117(42):12887–12897.
- [Peisajovich and Tawfik, 2007] Peisajovich, S. G. and Tawfik, D. S. (2007). Protein engineers turned evolutionists. *Nat. Methods*, 4(12):991–994.
- [Street and Barrick, 2009] Street, T. O. and Barrick, D. (2009). Predicting repeat protein folding kinetics from an experimentally determined folding energy landscape. *Protein Science*, 18(1):58–68.
- [Sulkowska et al., 2012] Sulkowska, J. I., Morcos, F., Weigt, M., Hwa, T., and Onuchic, J. N. (2012). Genomics-aided structure prediction. *Proc. Natl. Acad. Sci. U.S.A.*, 109(26):10340–10345.
- [Tamaskovic et al., 2012] Tamaskovic, R., Simon, M., Stefan, N., Schwill, M., and Plückthun, A. (2012). Designed ankyrin repeat proteins (darpins) from research to therapy. *Methods Enzymol*, 503:101–134.
- [Weigt et al., 2009] Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. (2009). Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72.
- [Wetlaufer, 1973] Wetlaufer, D. B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 70(3):697–701.
- [Wetzel et al., 2008] Wetzel, S. K., Settanni, G., Kenig, M., Binz, H. K., and Plückthun, A. (2008). Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. *Journal of molecular biology*, 376(1):241–257.
- [Wolynes, 1996] Wolynes, P. G. (1996). Symmetry and the energy landscapes of biomolecules. *Proceedings of the National Academy of Sciences of the United States of America*, 93(25):14249.

Supplementary material

MI and DI calculations over repeat protein families.



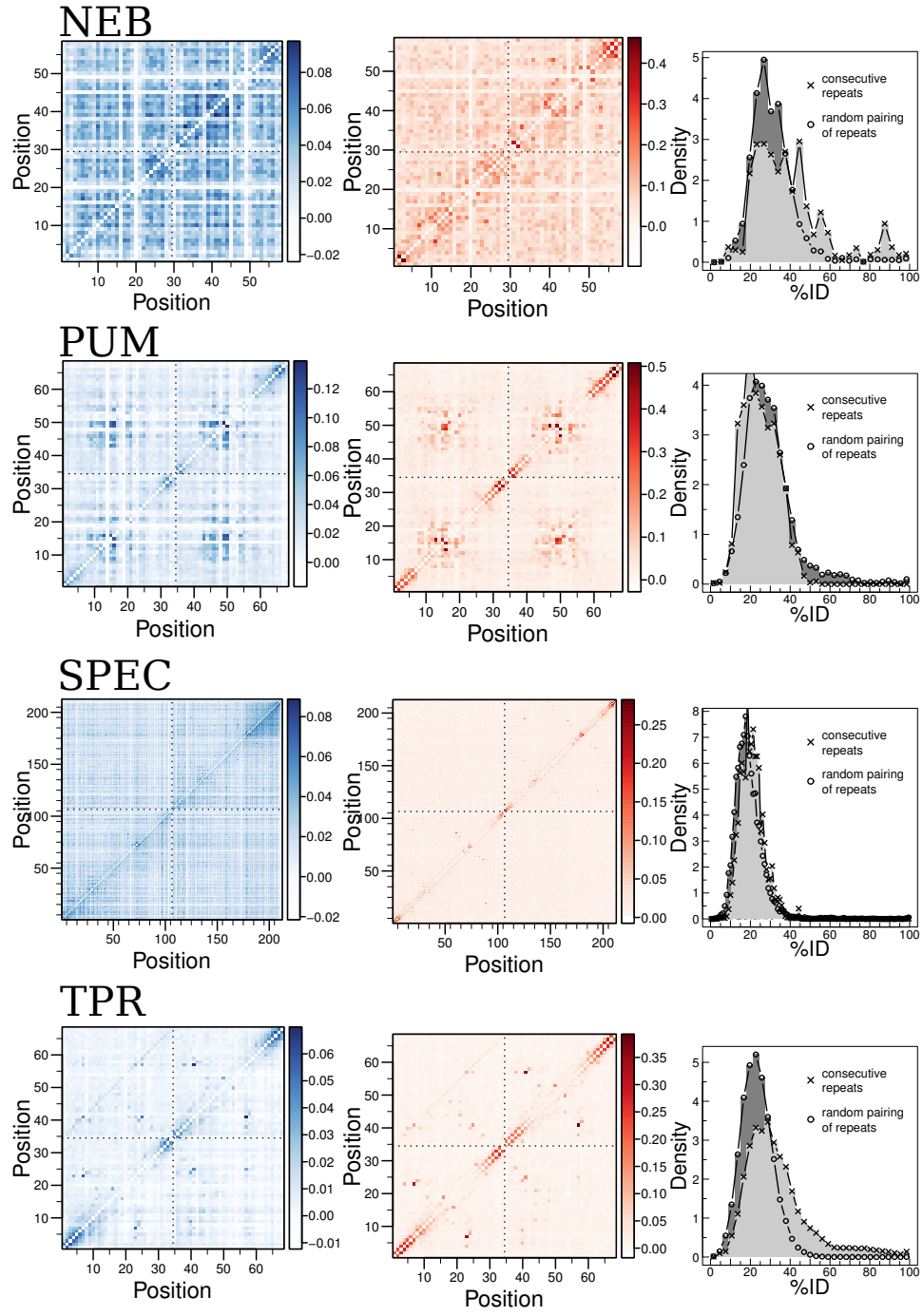
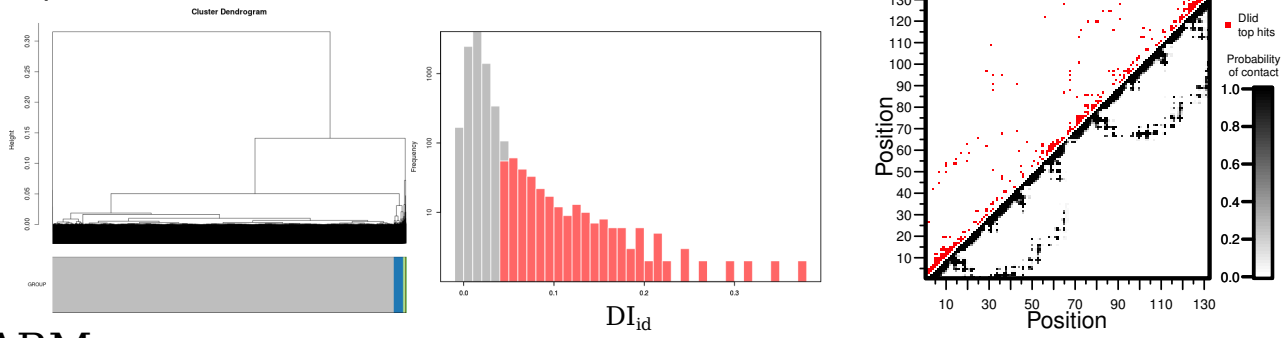


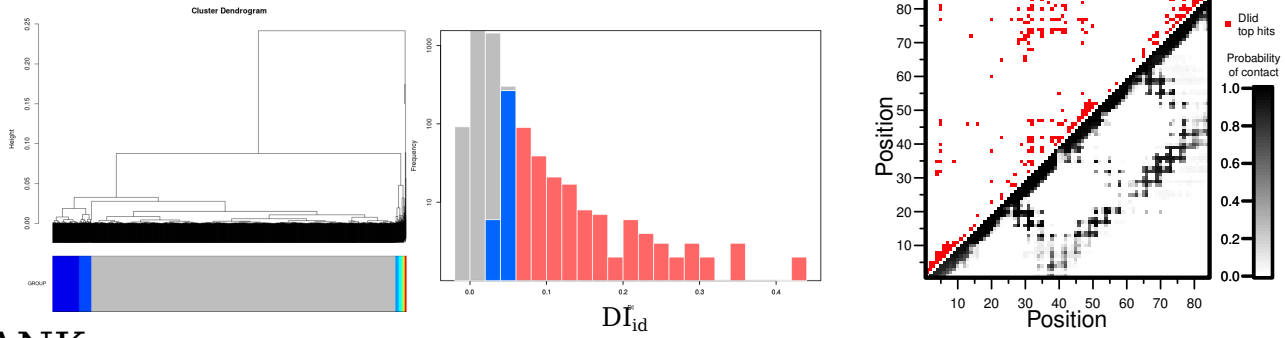
Figure S1: For each family, the blue matrix is MI on the upper triangle and MI_{id} on the lower triangle; the red matrix is DI on the upper triangle and DI_{id} on the lower triangle; the third panel has the comparison between histograms of %ID for the FNA (first neighbours repeats - x) and the RPA (random pairs of repeats alignment - o).

Selection of top DI hits.

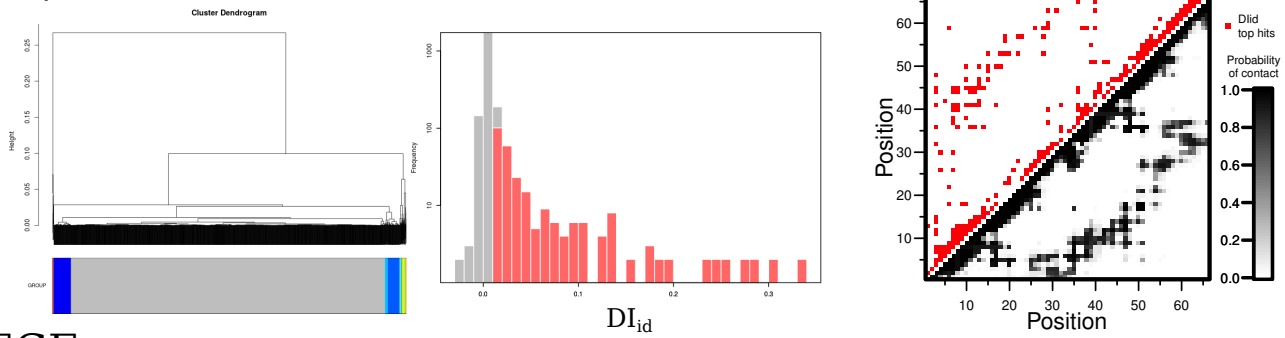
ANEX



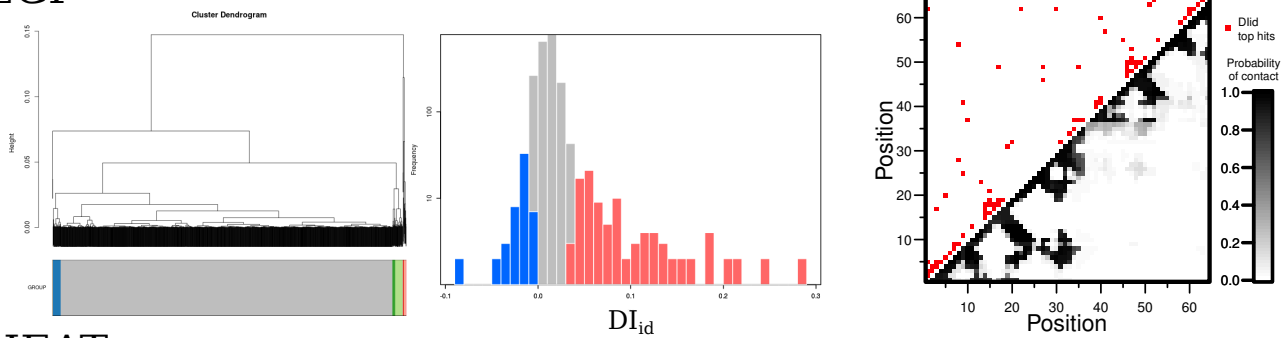
ARM



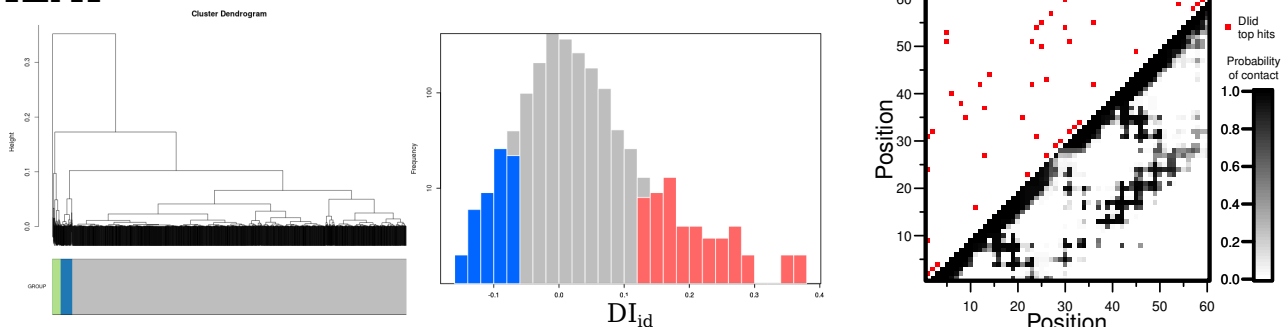
ANK



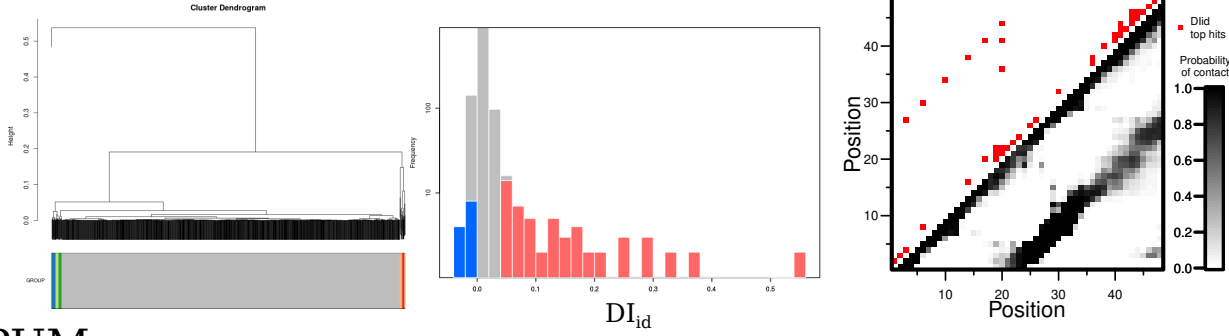
EGF



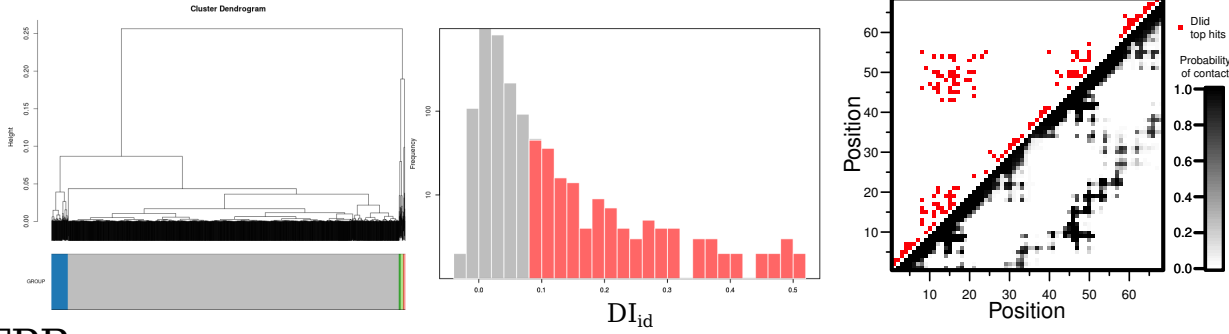
HEAT



LEU



PUM



TPR

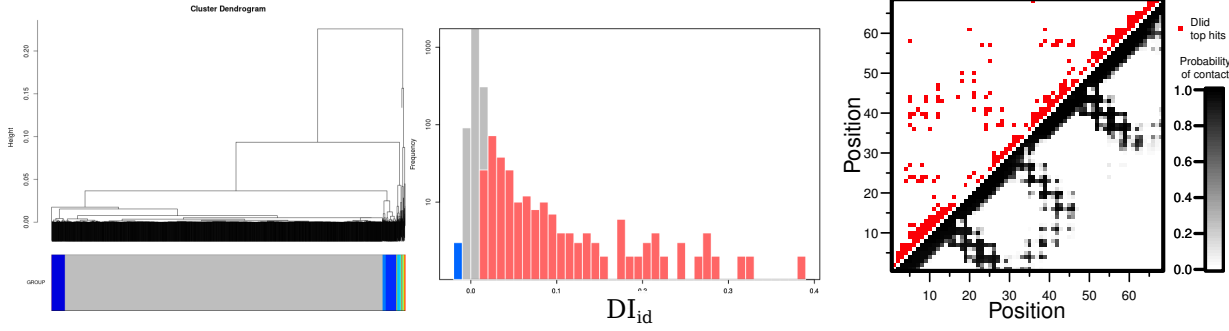


Figure S2: For each protein family, on the left the dendrogram of dDI_{id} : each leaf is a pair of positions and the height is the absolute value of difference of DI. In the center the histogram of DI_{id} values. On red and blue the small clusters distributions; on red ones considered DI_{id} hits, and on blue the ones that were not. The third panel, on top the DI_{id} hits and below the probability of contact map.

Distant couplings along a repeat-array

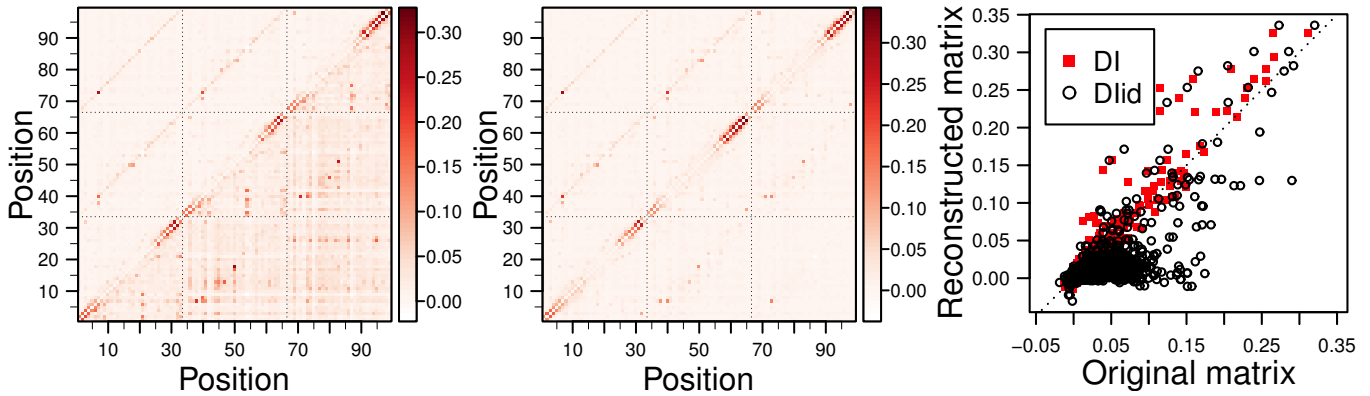


Figure S3: Left, upper triangle DI and bottom triangle DI_{id} for the three repeats alignment. Center, upper triangle DI and bottom triangle DI_{id} calculated from different alignments (first neighbours and second neighbours pairs) and reconstructing the matrix. Right, comparison of the DI and DI_{id} values obtained on the first two panels.

Table 1: Repeat protein families analyzed. L is the number of residues of the sequences on the FNA; M is the number of sequences on the FNA.

Family name	Abbreviation	Pfam Identifier	$2L_0$	M
ANKYRIN	ANK	PF00023	66	72908
TETRATRICOPEPTIDE	TPR	PF00515	68	38866
LEUCINE RICH	LEU	PF00560	48	26493
SPECTRIN	SPEC	PF00435	212	13142
EPIDERMAL GROWTH FACTOR	EGF	PF00008	64	10842
ARMADILLO	ARM	PF00514	84	6911
ANNEXIN	ANEX	PF00191	132	4264
PUMILIO	PUM	PF00806	68	3995
NEBULIN	NEB	PF00880	58	2438
HEAT	HEAT	PF02985	60	261

Robustness and confidence of the analysis

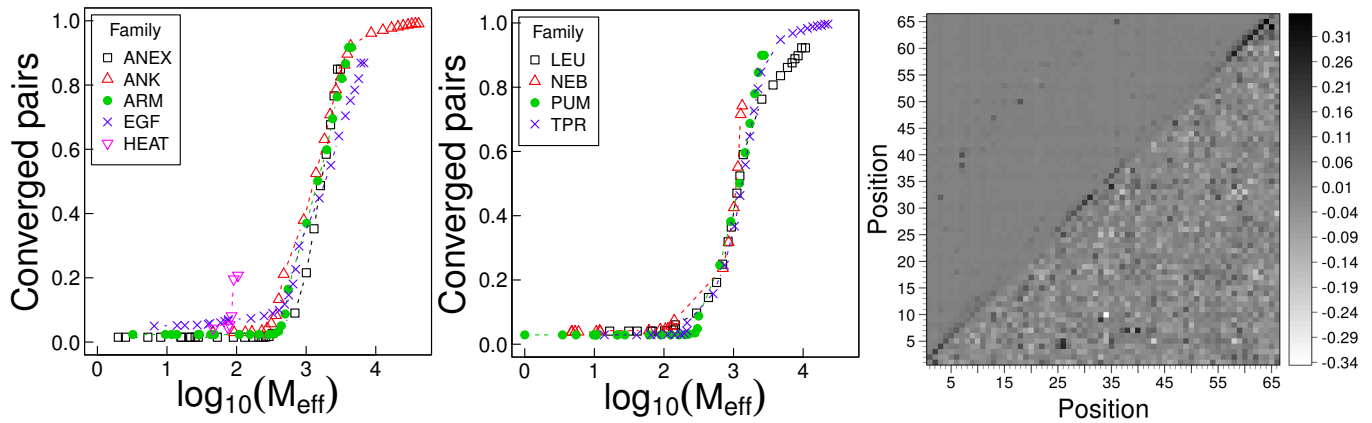


Figure S4: Left and center, for each family proportion of pairs of positions converged according to the criteria of the main text vs. the number of effective sequences on the alignment. Right, for the ANK family, example of $DI_{i,d}$ matrix calculated over an alignment of around 70000 sequences (upper panel) and over an alignment of around 400 sequences (lower panel).